The Avalon Beowulf Cluster

Michael S. Warren

Theoretical Astrophysics, (T-6)
Los Alamos National Laboratory
Los Alamos, NM

Software Engineering Department
Digital Island, Inc.
Thousand Oaks, CA

Chairman of the Board of Directors
Beomax, Inc.
Pismo Beach, CA

Supercomputing ’96

Avalon

Avalon Timeline

May 16, 1996	LDRD proposal for Beowulf research rejected
Aug. 13, 1996	Memo to T director proposing Beowulf cluster
Sep. 30, 1996	Loki cluster built with T-div funding
Nov. 20, 1997	Loki wins Bell price/performance prize
Feb. 2, 1998	Meeting at CNLS to discuss 64 node Alpha cluster
Mar. 10, 1998	Machines ordered from Carrera
Apr. 10, 1998	70 nodes delivered
Apr. 13, 1998	Machine fully assembled, Linpack over 10 Gflops
Jun. 18, 1998	Avalon ranked 315 on Top500 list
Jun. 25, 1998	First node failure in 47 days
Sep. 11, 1998	Upgrade to 140 nodes completed
Nov. 5, 1998	Avalon ranked number 113 on Top500 list
Nov. 12, 1998	Avalon wins Bell price/performance prize
Nov. 11, 1999	Avalon falls to number 265 on Top500 list
Nov. 2000	Avalon falls off the Top500 list
July 2001	3 million node hours provided, 60+ publications

Avalon

Loki

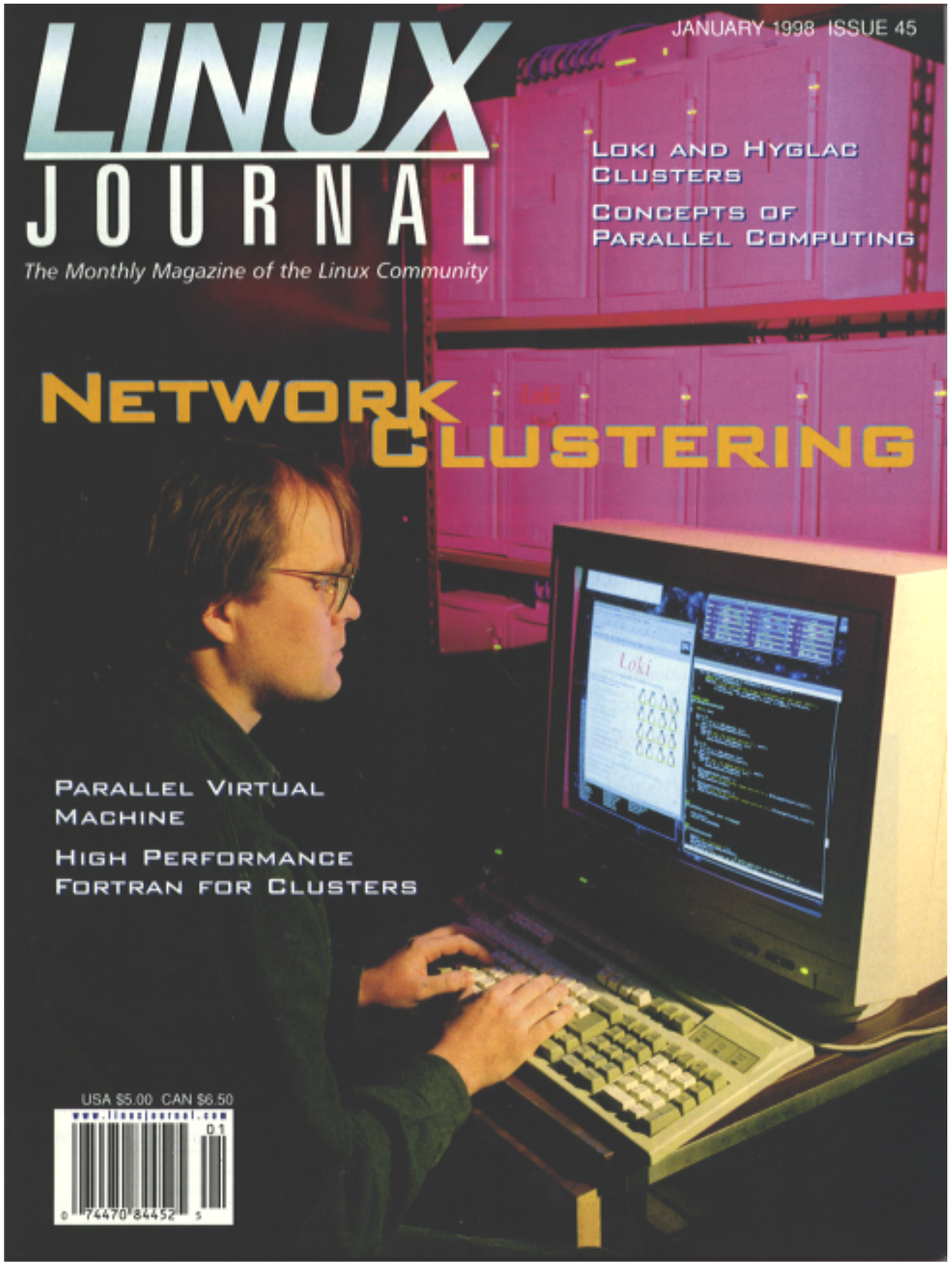
Qty.	Price	Ext.	Description
16	595	9520	Intel Pentium Pro 200 Mhz CPU/256k cache
16	15	240	Heat Sink and Fan
16	295	4720	Intel VS440FX (Venus) motherboard
64	235	15040	8x36 60ns parity FPM SIMMS (128 Mb per node)
16	359	5744	Quantum Fireball 3240 Mbyte IDE Hard Drive
16	85	1360	D-Link DFE-500TX 100 Mb Fast Ethernet PCI Card
16	129	2064	SMC EtherPower 10/100 Fast Ethernet PCI Card
16	59	944	S3 Trio-64 1Mb PCI Video Card
16	119	1904	ATX Case
2	4794	9588	3Com SuperStack II Switch 3000, 8-port Fast Ethernet
		255	Ethernet cables
Total		$51,379	

Table 1: Loki architecture and price (September, 1996).

Avalon

Avalon Architecture (After Upgrade)

Qty.	Price	Ext.	Description
140	1701	238140	DEC Alpha 164LX 533 MHz 21164A, with 2x128Mb SDRAM DIMM ECC memory (256 Mbyte/node), Quantum 3240 Mbyte IDE Hard Drive, Kingston 100 Mb Fast Ethernet PCI Card, cables, assembly, Linux install, 3 year parts/labor warranty
70	285	19950	128 Mb Memory upgrade for initial 70 nodes
4	6027	24108	3Com SuperStack II 3900, 36-port Fast Ethernet
8	968	7744	Gigabit uplink modules for 3900s
1	10046	10046	3Com SuperStack II 9300, 12-port Gigabit Ethernet
5	1055	5275	Cyclades Cyclom 32-YeP serial concentrators
140	10	1400	Serial cables (20 ft)
7	117	819	Shelving
56	100	5600	Final assembly labor
Total		$313,082	$2236 per node 1.066 Gflops peak per node


LINUX
JOURNAL
The Monthly Magazine of the Linux Community
JANUARY 1998 ISSUE 45
Loki and Hyglac Clusters
Concepts of Parallel Computing
Network Clustering
Parallel Virtual Machine
High Performance Fortran for Clusters
USA $5.00 CAN $6.50
www.linuxjournal.com
D1
0 74470 84452 5

1997 GORDON BELL PRIZE

Winner
Performance

Michael S. Warren
Los Alamos National Laboratory

John K. Salmon
California Institute of Technology

Winner
Price/Performance

Michael S. Warren and M. Patrick Goda
Los Alamos National Laboratory

Donald J. Becker
NASA Goddard Space Flight Center

John K. Salmon and Thomas Sterling
California Institute of Technology

Grégoire S. Winckelmans
Catholic University of Louvain

In recognition of their superior effort in practical parallel-processing research

Barry W. Johnson

Barry W. Johnson, President, IEEE Computer Society

Ed Parrish, Editor-in-Chief, Computer

IEEE COMPUTER SOCIETY IEEE

Presented by COMPUTER

Avalon

1998 GORDON BELL PRIZE

Second Prize
Performance per Dollar Category

Michael S. Warren, Timothy C. Germann, Peter S. Lomdahl, David M. Beazley
(Los Alamos National Lab.)

John K. Salmon (California Inst. of Technology)

For Their Superior Efforts in
Avalon: An Alpha/Linux Cluster Achieves 10 Gflops/s for $150K

IEEE COMPUTER SOCIETY

IEEE

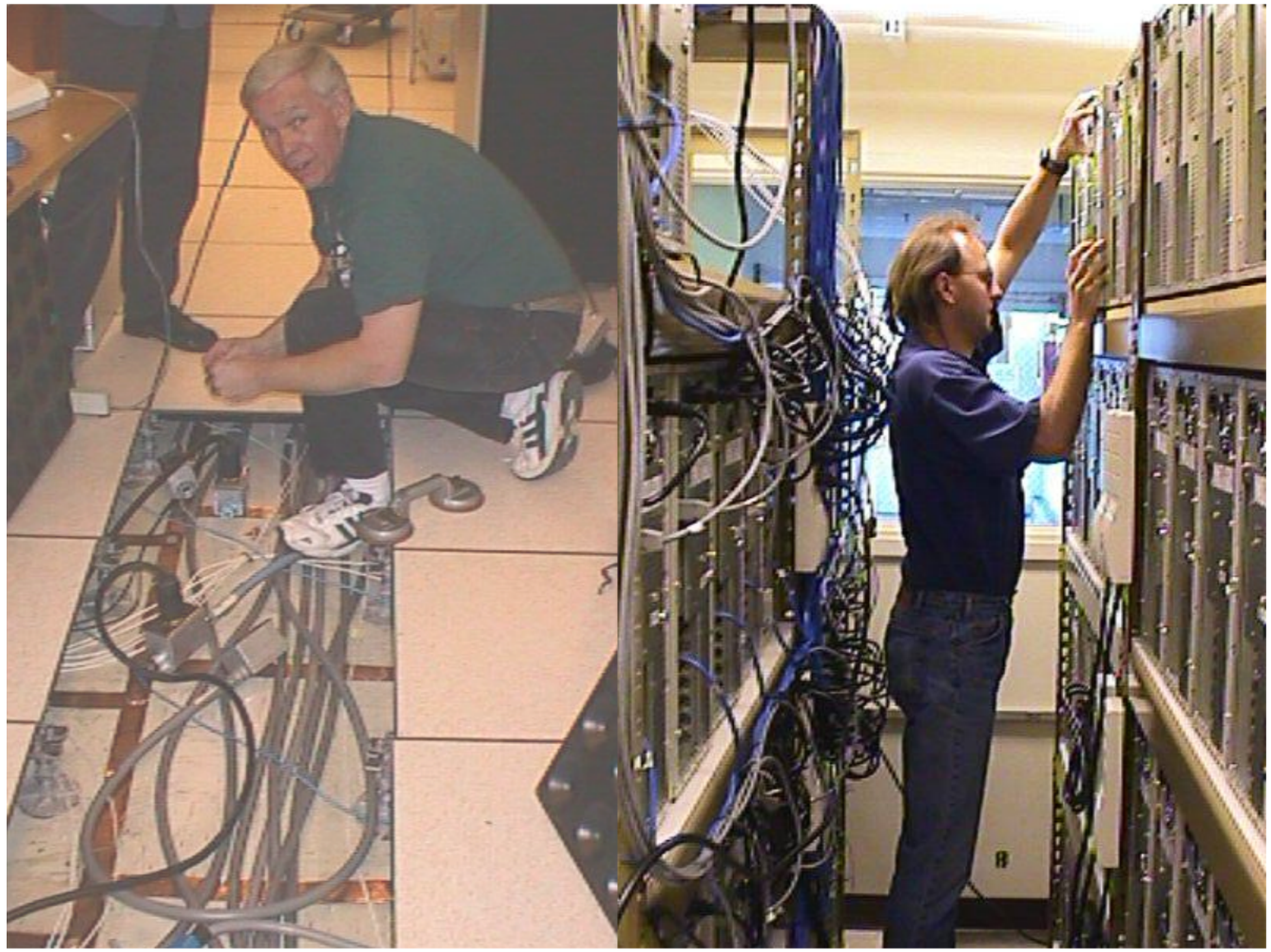
Doris L. Carver, President, IEEE Computer Society

Avalon

Avalon Construction

Avalon

Avalon Construction



Avalon

Avalon Construction

Avalon

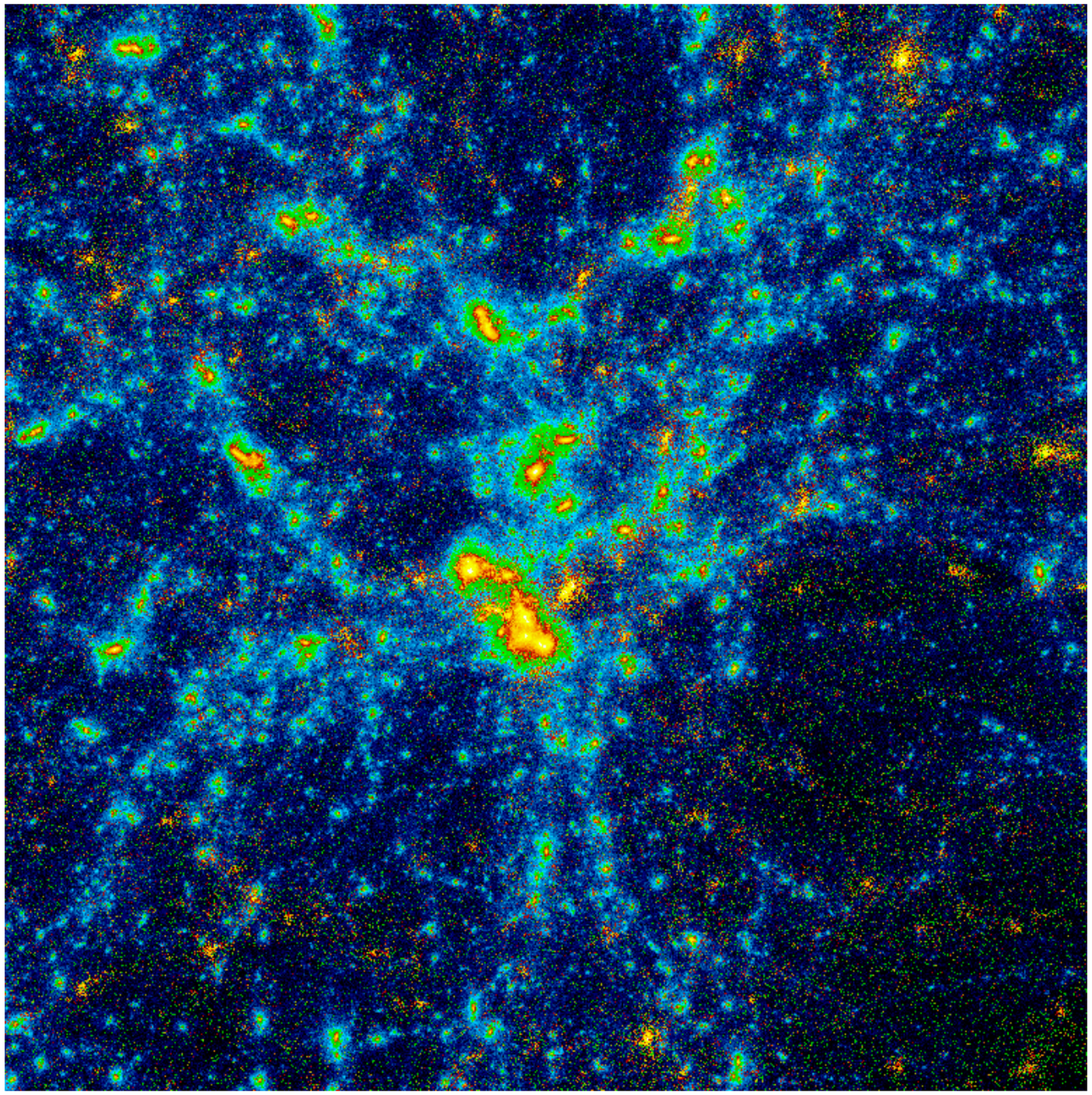
Avalon

Avalon Simulations

- In April 1998, Avalon performed a 60 million particle molecular dynamics (MD) simulation of shock-induced plasticity using the SPaSM MD code. This simulation ran for a total of 332 hours on Avalon, computing a total of 1.12×10^{16} floating point operations.
- The beginning of this simulation sustained 9.9 Gflops over a 44 hour period, and saved 68 Gbytes of raw data.
- Overall, we obtained a sustained throughput of 9.4 Gflops and a price/performance of $16/Mflop (or $12/Mflop without checkpoints or graphics).

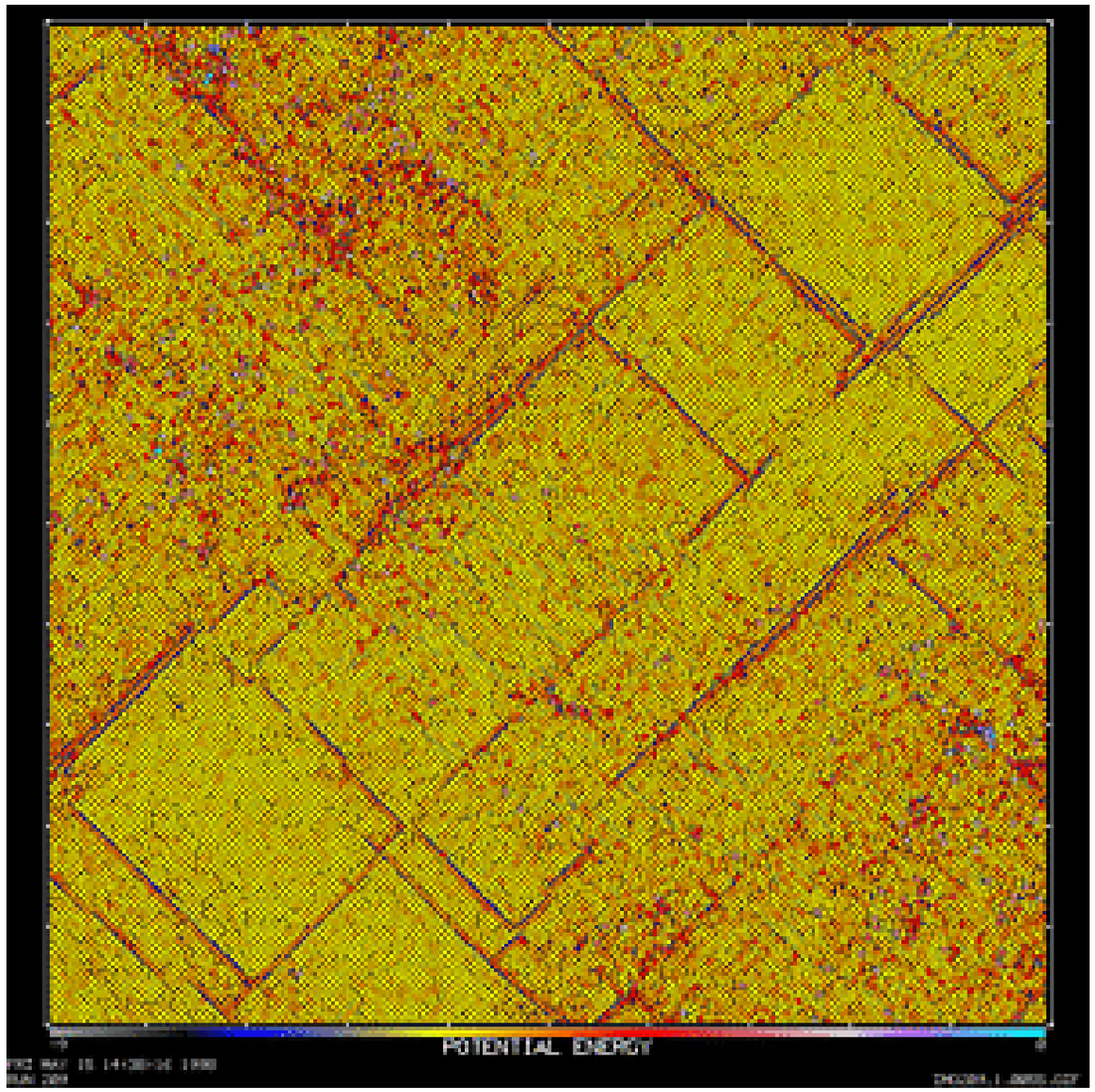
Avalon

Galaxy Formation Simulation



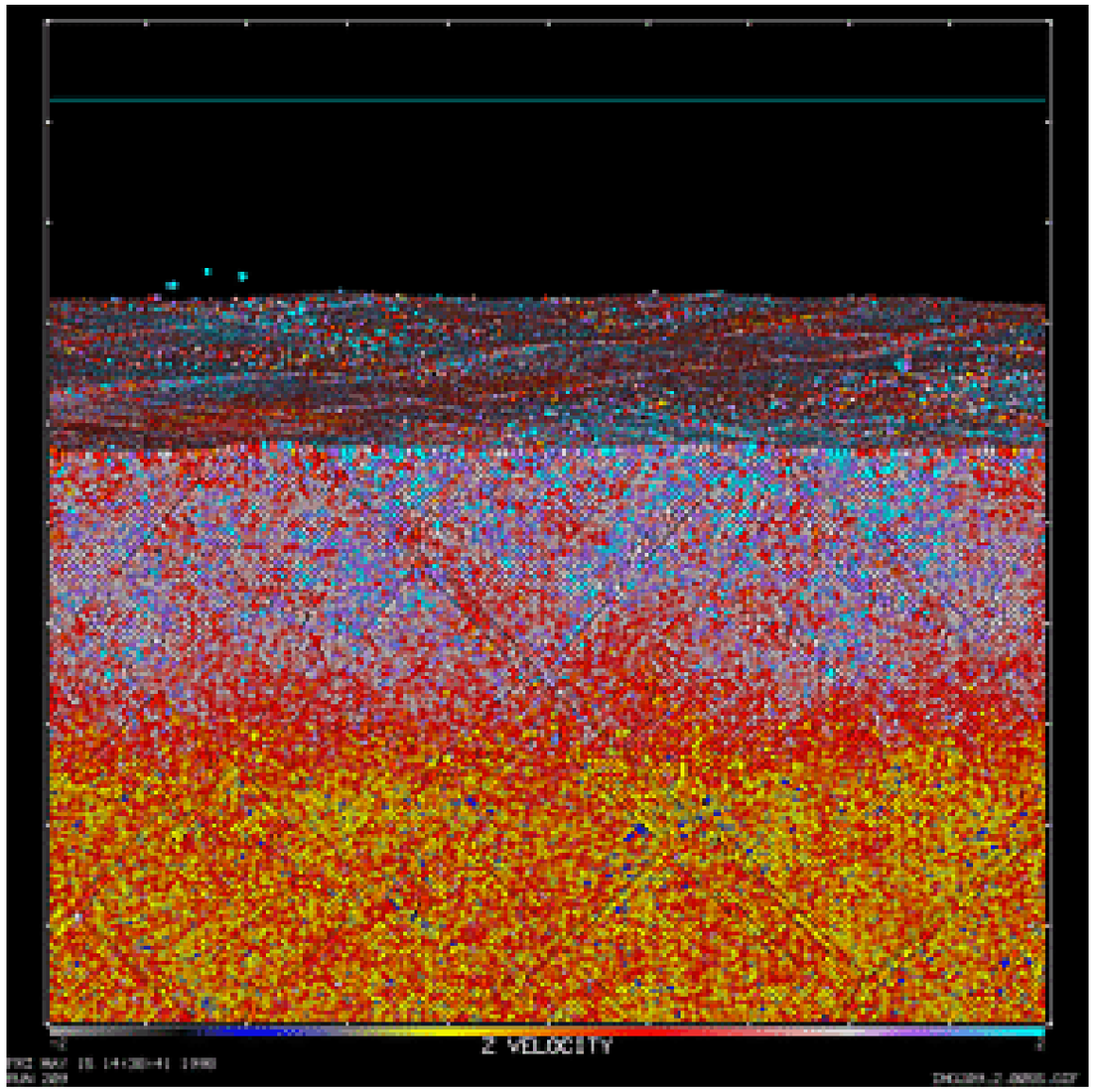
Avalon

SPaSM Molecular Dynamics Simulation



Avalon

SPaSM Molecular Dynamics Simulation



Avalon

Avalon Simulations

- In April 1998, Avalon performed a gravitational treecode N-body simulation of galaxy formation using 9.75 million particles, which sustained an average of 6.78 Gflops over a 26 hour period.
- This simulation is exactly the same as that which won a Gordon Bell price/performance prize in 1997 on the Loki cluster, at a total performance 7.7 times that of Loki, and a price/performance 2.6 times better than Loki.
- Avalon was upgraded to 140 processors and 35 Gbytes of memory in September, and was able to update 150 million particles at 29.6 Gflops, resulting in a price/performance of $10.5/Mflop.

Avalon

Looking back

Year	Machine	Perf.	Price/Perf	Total flops
1998	Avalon	10 Gflops	$16	1.1×10^{16}
1997	Loki	880 Mflops	$58	1.2×10^{15}
1996	6 proc SGI R8000	903 Mflops	$160	3.8×10^{11}
1995	20 proc HP 9000	176 Mflops	$450	4.6×10^{13}
1994	8 proc HP 9000	780 Mflops	$300	3.9×10^{9}
1993	SNAP-32	409 Mflops	$133	1.5×10^{8}

Table 2: Past price/performance statistics.

The total number of floating point operations performed in the Avalon MD simulation was greater than the number of operations carried out on ASCI Red by last year's Gordon Bell Performance Prize winner.

Avalon

Treecode performance

Site	Machine	Procs	N $\times 10^6$	Gflops	Mflops/proc
Sandia	ASCI Red	6800	322	464.9	68.4
LANL	Avalon	128	150	17.60	137
Sandia	ASCI Red	4096	10	164.3	40.1
LANL	Avalon	128		16.16	126
LANL	TMC CM-5	512		14.06	27.5
Caltech	Intel Paragon	512		13.70	26.8
LANL	SGI Origin 2000	64		13.10	205.0
NRL	TMC CM-5E	256		11.57	45.2
Caltech	Intel Delta	512		10.02	19.6
NAS	IBM SP-2(66/W)	128		9.52	74.4
JPL	Cray T3D	256		7.94	31.0
Caltech	Naegling	96		5.67	59.1
LANL	CM-5 no vu	256		2.62	5.1
SC ’96	Loki+Hyglac	32		2.19	68.4
LANL	Loki	16		1.28	80.0

Avalon

Production Computing

- Avalon currently provides over 18,000 node-hours of production computing time per week, split among about 10 production users.
- Avalon provides 4000 node-hours of development time per week for another 40 users.
- Obtaining an equivalent amount of computing through Los Alamos institutional sources would cost a minimum of $40,000 per week ($2 million/yr).

Avalon

Real Costs

- Initial hardware cost of about $300k. Initial software cost of $0.
- Power and space for the machine are estimated to be about $20k/yr.
- All of the hardware and software maintenance on the machine is performed in the spare time of four people, averaging less than 10 man-hours of labor per week ($50k/yr at $100/hr).
- Amortized over three years, the real cost (including all incidentals) works out to about 15 cents per cpu hour.

Avalon

TOP500 supercomputer SITES

- Avalon ranked at #113 on the TOP500 supercomputers list in Nov. 1998, at 48,600 Mflops.
- TOP500 is based on Parallel Linpack performance.
- Avalon peak performance is 149,400 Mflops.
- Falls between a 156 processor IBM SP PC604e 332 and a 25 processor NEC SX-4
- In March 2000 Avalon ranked at #265 on the TOP500 list. Currently (July 2001), the end of the top500 list is at 68 Gflops.

Avalon

Microsoft's words from
"Linux OS Competitive Analysis"

Beowulf clustering - Beowulf is a shared-nothing cluster that runs today on Linux. It requires specially developed applications which are able to spawn subprocesses on remote hosts for computing. As such, it is not a real competitor to WolfPack and most of the magic in Beowulf [is] in the applications rather than system services. However, as a press-magnet, Beowulf clusters with appropriate software have been demonstrated at supercomputer power (a 10GFLOP was recently ranked #315 on the top 500 supercomputers list maintained by the NCSA).

Avalon

Linux vs. NT, from Microsoft's memo

Real or perceived advantages.

- Customization - The endless customizability of Linux for specific tasks - ranging from GFLOP clustered workstations to 500K RAM installations to dedicated, in-the-closet 486-based DNS servers - makes Linux a very natural choice for "isolated, single-task" servers such as DNS, File, Mail, Web, etc. Strict application and OS componentization coupled with readily exposed internals make Linux ideal.

- Availability/Reliability There are hundreds of stories on the web of Linux installations that have been in continuous production for over a year. Stability more than almost any other feature is the #1 goal of the Linux development community (and the #1 cited weakness of Windows)

Avalon

Linux vs. NT, from Microsoft's memo

Real or perceived advantages.

- Scalability/Performance Linux is considered faster than NT in networking, and processes. In particular, as a server, Linux's modular architecture allows the administrator to turn off graphics, and other non-related subsystems for extreme performance in a particular service.

- Interoperability Every open protocol on the planet (and many of the closed ones) have been ported to Linux. In a Windows environment, work from the SAMBA team enables Linux to look like an NT Domain Controller / File Server.

products

ALPHASERVER

compaq.com/products

Linux–Ready AlphaServer Systems

AlphaServer Home

AlphaServer Linux tested options list

Compaq Linux website

"Linux–ready" AlphaServer systems
AlphaServer DS20
AlphaServer 800

US Resellers

Digital Domain studios used the winning combination of Alpha systems and Red Hat Linux to create virtually every frame of the Academy award–winning movie Titanic.

The Los Alamos National Laboratory used Alpha–powered Linux systems to create "Avalon," one of the world's fastest supercomputers (113 out of 500) –– at a "dirt–cheap" price.

AlphaServer systems are ideally suited to run Linux –– and run it faster and better than anyone else. Many companies are looking to Linux to solve difficult computing problems in a user–friendly, stable environment. And others see Linux as a cost–effective way to begin building for a future that will include Tru64 UNIX.

The Compaq AlphaServer DS20 and 800 systems are available without a bundled software license, allowing you to use an open source operating system, such as Linux. This saves you the cost of purchasing an operating system you don't plan to use.

Linux is a freely available operating system that was started in 1991 by Linus Torvalds in Helsinki Finland. It has grown over the last 8 years to boast an installed base of 7–10 million users. According to IDC the Linux server shipments grew over 212% from 1997 to 1998 from 1/4 million to 3/4 of a million making Linux one of the top growing operating systems for servers.

For competitive pricing and more information, use this form and we will have one of our partners in your local area contact you. US part numbers and list prices are as follows:

Part Number	Description	U.S. List Price
DJ–55NJA–CA	AlphaServer DS20 6/500, Pedestal, 128 MB, 4 MB cache, no software	16,870
DJ–55NJA–DA	AlphaServer DS20 6/500, Pedestal, 256MB, 4 MB cache, no software	18,073
DJ–55NJA–EA	AlphaServer DS20 6/500, Pedestal, 512MB, 4 MB cache, no software	20,992
3X–PB82B–XA	AlphaServer 800 5/500, Pedestal, 64MB memory, 2 MB cache, 4.3 GB UltraSCSI disk	5,690
3X–PB82P–XA	AlphaServer 800 5/500, Rackmount, 64MB memory, 2 MB cache, 4.3 GB UltraSCSI disk	5,940

***Note**: The systems come with a three year on–site hardware warranty from Compaq directly. The resellers, by arrangement with selected LINUX software providers, will support the software warranty.*

Avalon

Avalon

In Conclusion

- The Beowulf architecture provides a low-cost message-passing hardware and software environment capable of performing large-scale numerical simulations.
- We hope the example we have provided with Avalon will encourage others to investigate the applicability of Beowulf-class hardware to other scientific supercomputing problems, as well as to other data-intensive computing projects.
- A good algorithm can achieve an increase of many orders of magnitude in performance, while hardware is limited to a mere factor of ten every five years.

Avalon

References

[1] N. D. Antunes, L. M. A. Bettencourt, and M. Kunz. The thermodynamics of monopoles in O(3) scalar field theory. *Phys. Rev. D*, 2000. (to be submitted).

[2] N. D. Antunes, L. M. A. Bettencourt, and A. Yates. Predicting the critical density of defects in O(N) scalar field theories. *Phys. Rev. Lett.*, 2000. (submitted).

[3] N. D. Antunes, L. M. A. Bettencourt, and W. H. Zurek. Vortex string formation in a 3D U(1) temperature quench. *Phys. Rev. Lett.*, 82:2824–2827, 1999.

[4] D. Bedrov, G. D. Smith, and T. D. Sewell. Pressure dependent shear viscosity coefficient of HMX. A molecular dynamics simulation study. (in preparation), 2000.

[5] D. Bedrov, G. D. Smith, and T. D. Sewell. Temperature dependent shear viscosity coefficient of HMX. A molecular dynamics simulation study. *J. Chem. Phys.*, 2000. (in press).

[6] E. Ben-Naim, S. Y. Chen, G. D. Doolen, , and S. Redner. Shocklike dynamics of inelastic gases. *Phys. Rev. Lett.*, 83:4069, 1999.

[7] L. M. A. Bettencourt, N. D. Antunes, and W. H. Zurek. The Ginzburg regime and its effects on topological defect formation. *Phys. Rev. D*, 2000. (submitted).

[8] L. M. A. Bettencourt, K. Rajagopal, and J. Steele. A model for simulating the chiral crossover in heavy ion collision experiments. *Phys. Rev. D*, 2000. (to be submitted).

[9] L. M. A. Bettencourt, R. Sasik, and S. Habib. Thermal motion of a single vortex in a bose-einstein condensate. *Phys. Rev. B*, 2000. (submitted).

[10] L. M. A. Bettencourt and W. H. Zurek. Vortex formation in a coherent velocity flow. *Phys. Rev. Lett.*, 2000. (to be submitted).

[11] A. Blotz. The spin structure of constituent quarks in the instanton liquid model. (in preparation), 2000.

[12] K. Camarda, P. Laguna, W. Miller, and M. S. Warren. Coalescence of a black hole with a neutron star and the formation of a dense accretion torus. *Ap. J.*, 2000. (submitted).

[13] S. Y. Chen, Y. Deng, X. Nie, and Y. Tu. Clustering kinetics of granular media in three dimensions. *Physics Letters A*, 2000. (in press).

[14] D. A. Egolf. Equilibrium regained: From nonequilibrium chaos to statistical mechanics. *Science*, 287:101–104, 2000.

[15] C. L. Fryer and M. S. Warren. Core-collapse supernovae in three-dimensions. *Ap. J.*, 2000. (in preparation).

[16] T. C. Germann and P. S. Lomdahl. Recent advances in large scale atomistic materials simulations. *Computing in Sci. and Eng.*, 1(2):10–11, 1999.

[17] J. Glimm, J. Grove, X. L. Li, W. Oh, and D. H. Sharp. A critical analysis of Rayleigh-Taylor growth rates. *J. Comp. Phys.*, 1999. (submitted).

[18] J. Glimm, S. Hou, H. Kim, D. H. Sharp, K. Ye, and Q. Zou. Risk management for petroleum reservoir production: A simulation-based study of prediction with confidence intervals. (in preparation), 1999.

[19] M. P. Goda, J. G. Hills, and M. S. Warren. Tsunami hazard from asteroid impact. (in preparation), 2000.

[20] J. E. Gubernatis and N. Hatano. The multicanonical monte carlo method. *Computer Simulations*, 2000. (to appear).

[21] N. Hatano and J. E. Gubernatis. Bivariate multicanonical monte carlo of the 3d $\pm j$ spin glass. In M. Tokuyama and I. Oppenheim, editors, *Slow Dynamics in Complex Systems*, pages 565–566, Maryland, 1999. AIP.

[22] N. Hatano and J. E. Gubernatis. A bivariate multicanonical monte carlo of the 3d $\pm j$ spin glass. In D.P. Landau, editor, *Recent Developments in Computer Simulation Studies in Condensed Matter Physics*, pages 149–161, Berlin, 2000. Springer.

[23] N. Hatano and J. E. Gubernatis. Evidence for the droplet picture of the 3d $\pm j$ spin glass. (in preparation), 2000.

[24] N. Hatano and J. E. Gubernatis. A multicanonical monte carlo of the 3d $\pm j$ spin glass. *Prog. Theor. Phys. Suppl.*, 2000. (to appear).

[25] B. L. Holian, T. C. Germann, P. S. Lomdahl, J. E. Hammerberg, and R. Ravelo. Shock waves and their aftermath: a view from the atomic scale. In M. D. Furnish, editor, *Shock Compression of Condensed Matter – 1999*, 2000.

[26] K. Kadau, T. C. Germann, P. S. Lomdahl, and B. L. Holian. Microscopic view of structural phase transitions due to shock waves. *Science*, 2000. (to be submitted).

[27] Z. Karkuszewski, C. Jarzynski, and W. Zurek. Decoherence due to chaotic environment: Structure saturation and the absence of the quantum butterfly effect. (in preparation), 2000.

[28] P. Möller and A. Iwamoto. Realistic fission saddle-point shapes. *Phys. Rev. C*, 61, 2000.

[29] P. Möller, D. G. Madland, and A. Iwamoto. Calculation of onset of fission mass asymmetry based on 5-dimensional 2 500 000 grid-point potential-energy surfaces. In *Tenth International Symposium on Capture Gamma-Ray Spectroscopy and Related Topics*. World Scientific, 1999.

[30] X. Nie, S. Y. Chen, and E. Ben-Naim. Dynamics of vibrated granular monolayer. *Europhysics Letters*, 2000. (in press).

[31] X. Nie, G. D. Doolen, and S. Y. Chen. Lattice-boltzmann simulations of fluid flows in MEMS. *Physics of Fluids*, 1999. (submitted).

[32] B. Stojkovic. Dynamics of sliding charge density waves. (in preparation), 2000.

[33] B. Stojkovic. Size effects in the model glass transition. (in preparation), 2000.

[34] T. C. Wallstrom, S. Hou, M. A. Christie, L. J. Durlofsky, D. H. Sharp, and Q. Zou. Effective medium boundary conditions for upscaling relative permeabilities. *Transport in Porous Media*, 1999. (submitted).

[35] M. S. Warren, T. C. Germann, P. S. Lomdahl, D. M. Beazley, and J. K. Salmon. Avalon: An Alpha/Linux cluster achieves 10 Gflops for $150k. In *Supercomputing '98*, Los Alamitos, 1998. IEEE Comp. Soc.

[36] M. S. Warren, A. A. Hagberg, and J. D. Moulton. A cartesian tree-based parallel adaptive partial differential equation solver. (in preparation), 2000.

[37] M. S. Warren, A. A. Hagberg, J. D. Moulton, D. Neal, and J. K. Salmon. Avalon: Champagne computing on a beer budget. (extended abstract), 1999.

[38] M. S. Warren and W. H. Zurek. The Los Alamos synthetic sky survey. *Ap. J.*, 2000. (in preparation).

[39] S. M Zoldi, V. Ruban, A. Zenchuk, and S. Burtsev. Parallel implementations of the split-step fourier method for solving nonlinear schrodinger type systems. *SIAM News*, 32, 1999.

[40] W. H. Zurek, L. M. A. Bettencourt, J. Dziarmaga, and N. D. Antunes. Topological defects and the non-equilibrium dynamics of symmetry-breaking phase transitions. In H. Godfrin and Y. Bunkov, editors, *Les Houches Lectures of NATO A.S.I.*, 2000.